Wrangle Report

By Frederick Yen

Overview

The data wrangling is performed for Udacity Data Analysis Nanodegree program Wrangle-and-Analyze project. The wrangled dataset is the tweet archive of WeRateDogs.

Wrangling Procedures

1.Gathering data

Three datasets were gathered from provided sources:

1. Twitter archive file, twitter_archive_enhanced.csv, was downloaded from link in project descriptions.
2. Tweet image predictions was downloaded programmatically using the Requests library and URL in project descriptions.
3. Twitter API and the Tweepy library was used to access the WeRateDogs Twitter archive. The tweet JSON data was queried and stored to tweet_json.txt file. The JSONs are read and stored into a dataframe in Jupyter notebook.

2. Assessing data

1. Visually assessment: Print contents of the dataframes.
2. Programmatic assessment: Inspect with pandas commands such as sample, info, value_counts, sample, duplicated, isnull, etc.
3. Eight **unique** quality issues and two **unique** tidiness issues were defined at this step. The numbers assigned for each bullet point indicates which unique issue it's assigned to.

Quality

- **archived**
    - (1) Remove unneeded columns
    - (2) Remove RTs and original tweets that don't have images
    - (3) Timestamps column converted to date time objects
    - (4) Remove rows that have no ratings
    - (5) Fix numerators that contain decimals to float data type

    - (7) Correct denominators to float data type and set to 10
    - (8) Set invalid dog names to NaN
- **predictions**
    - 1) Remove unneeded columns
    - (6) Remove duplicated jpg URLs
- **tweet_json**
    - (2) Remove RTs

Tidiness

1. Tweet_id should have uniform data types between all tables
2. All tables should be merged to one master table

3. Cleaning data

The general steps of cleaning data are listed as follow:

1. Create a copy of the three original dataframes for cleaning.
2. Archived: Remove retweets and tweets without images
3. Archived: Change format of timestamps into formal date time objects
4. Archived: Manually clean out erroneous data entries with no ratings. This particular step was most time consuming due to the need for line-by-line manual inspection and then determining what/how needs to be fixed.
5. Archived:Corrected data type for both numerators and denominators. Numerators are then scaled accordingly as denominators are set to 10.
6. Archived: The replace operator function is called to replace all invalid names to nan.
7. For the predictions, a for loop was implemented to go through each row of the data frame, keeping the main prediction while removing other unneeded info.
8. For all dataframes, make sure to set tweet id to a uniformed datatype, to int. Lastly using the pandas merge function to merge dataframes into a master dataframe.